ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Impact of Social Media in Healthcare and Classification and Prediction of Swine Flu Related Tweets Using LDA Model

S.Mahalakshmi [1], Dr.M.Suriakala [2]

Research Scholar, Bharathiar University Coimbatore, India[1],

Professor, Dr.Ambedkar Govt. Arts College, Chennai, Tamilndu, India [2]

ABSTRACT: Today, a lot of and a lot of members of the medical community area unit mistreatment social media for sharing useful medical data and providing patient care. Social media is changing into a lot of and a lot of utilised by hospitals and medical professionals as a method to convey general health data, generally even personalised facilitate. Social media is gaining quality owing to its data spreading feature. Twitter is one in all the foremost powerful supply of data sharing as a result of its large users. Consequently, Twitter has become a preferred resource so as to the info for various research functions like social engineering, sentiment analysis, business functions etc. owing to its straightforward information availableness. In Twitter, the data that area unit re tweeted persistently are often treated as well-liked. During this analysis, we have a tendency to investigate the prediction of the recognition of messages by the amount of re- tweets and topic model is employed for extracting connected tweets for classification.

KEYWORDS: Information, LDA model, Classification, Twitter ,Social media.

I. Introduction

Social media is creating interactions between finish users and repair suppliers doable by providing comparatively straightforward, straightforward to access and unbiased platforms for sharing feedback. Several tending suppliers within the world area unit on social media like Twitter, Facebook, YouTube and blogs.

In this paper 1st tweets area unit retrieved from twitter.com by twitter API(training set), then supervisor learning technique referred to as Dirichlet Allocation (LDA)can be used and mistreatment this classifier .First we've to think about every tweet as a document and mistreatment LDA totally different topics area unit designated from the document so different words associated with the topics area unit retrieved and appointed to corresponding topics(classification) then we've to judge however well this classification is i.e., do the topics correspond well to the tweets and that we have to be compelled to predict the tweets that area unit well-liked.

This paper includes

Section II explains Review of connected Literature
Section III explains regarding Social Media
Section IV tending and Social Media
Section V flu
Section VI Machine Learning Algorithms
Section VII LDA Model , classification &amp; Prediction of tweets
Section VIII Conclusion.

II. RELATED WORK

Twitter has been used for many health-related functions, as well as to broadcast data regarding polygenic disorder [1], communicate throughout a disaster [2]and to know health-related trends and problems like flu [3], tobacco [4], downside drinking [5], dental pain [6], antibiotics and medicine misuse [7,8], et al. [9].Pang[10] think about word presence vs. frequency wherever word presence is found to be more practical than word frequency for sentiment analysis. Word position among a given sentence also can be effective, wherever such data are often accustomed decide if a specific word has a lot of strength at the start or the top of a given sentence. we have a tendency to expect that a broad vary of individuals can have the benefit of the results of the Social net program.

Social networks have compete a vital role in several domains for regarding one decade, notably concerned in an exceedingly broad vary of social activities like user interaction, establishing relationship relationships, sharing and recommending resources, suggesting friends, making teams and communities, commenting friend activities and opinions and then on. Recent years, has witnessed the fast progress within the study of social networks for numerous applications, like user identification in Face book and cluster recommendation via Flickr .Twitter[12] may be a social networking application that permits folks to small journal a few broad vary of topics. It helps users to attach with their followers. The tweets from users area unit noted as small blogs as a result of there's a one hundred forty character limit obligatory by Twitter fore very tweet. This lets the users gift any data with solely some words, optionally followed with a link to a lot of elaborated supply of data. The goal of our work is to mechanically classify incoming tweets into totally different classes so users aren't weak by the data.

Latent variable topic models are applied wide to issues in text modelling, and need no manually created coaching information. These models distill collections of text documents (here, tweets) into distributions of words that tend to co-occur in similar documents – these sets of connected words area unit noted as “topics”[11]. probabilistic topic models, a collection of algorithms that give a applied mathematics answer to the matter of managing giant archives of documents. With recent scientific advances in support of unattended machine learning versatile elements for modelling, scalable algorithms for posterior abstract thought, and increased access to large datasets topic models promise to be a vital part for summarizing and understanding our growing digitized archive of information[13]. As well-liked messages contain important data for the users, one needs to study the characteristics of such messages since it's associated with breaking news identification, infective agent selling and different similar tasks[14]. Given the large corpus of time period information being generated on a daily basis there has been associate degree increasing interest in Twitter information analysis, from modelling public sentiments to being associate degree indicator of poll results. it's been shown that despite their brevity, one hundred forty characters contains enough data to replicate political sentiments [15][16].

Recent years, tweets classification has become a preferred topic owing to the recognition of Twitter. Iranian in [17] planned a machine learning technique to mechanically establish trend-stuffing in tweets, mistreatment texts and links of tweets. Probabilistic topic models like LDA [19] area unit helpful for locating latent linguistics patterns from unstructured information in numerous forms like text, images, and user behaviour [20, 21, 22, 23, 24]. they're particularly convenient for user-generated contents on the net, like on-line news articles, blogs, and small blogs, wherever topics and different latent dimensions aren't well-known a priori. 2 sensible issues arise once applying topic models to net documents. First, a corpus of net documents is usually terribly giant, consisting of lots of documents, that makes posterior abstract thought quite slow. Second, a corpus of net documents expands chop-chop, therefore the model parameters should be updated incessantly to influence the new documents. Topic modelling is gaining progressively attention in several text mining communities. Latent Dirichlet Allocation (LDA) [25] is changing into a customary tool in topic modelling. As a result, LDA has been extended in an exceedingly form of ways in which, and particularly for social networks and social media, variety of extensions to LDA are planned.

III. About Social Media

Social media sites give a range of options that serve totally different functions for the individual user. they will embody blogs, social networks, video- and photo-sharing sites, wikis, or a myriad of different media, which may be classified by purpose, serving functions such as:

- Social networking (Facebook, MySpace, Google and, Twitter)
- Professional networking (LinkedIn)
- Media sharing (YouTube, Flickr)
- Content production (blogs [Tumblr, Blogger] and microblogs [Twitter])
- Knowledge/information aggregation (Wikipedia)
- Virtual reality and vice environments (Second Life)

Social media area unit computer-mediated tools that enable folks to form, share or exchange data, ideas, and pictures/videos in virtual and networks. Social media is outlined as "a cluster of Internet-based applications that depend upon the philosophic and technological foundations of net two.0, which enable the creation and exchange of user-generated content.

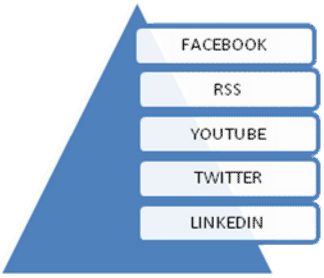

Fig 1.Social Media

Blogs also can give the chance to publish giant amounts of data in an exceedingly form of media (text, video, associate audio) in an open forum. Most blogging platforms enable readers to reply to printed content by posting their own comments. this allows associate degree in progress dialogue between the blogger and his or her audience. samples of wide used free “long-form” blogging platforms embody Tumblr (www.tumblr.com), WordPress (www.wordpress.org), and Blogger (www.blogger.com).

On Twitter, users publish messages (called “tweets”) that comprise a most of one hundred forty characters. Tweets are often supplemented with hyperlinks to different on-line media, like videos or websites. Tweets also can embody “hash tags,” a type of data compartmentalization that enables folks to go looking for tweets that area unit associated with a specific discussion or topic. Hash tags followed by HCPs embody #HCSM (for Health Care Social Media), #MDChat, and #Health20.

Media-sharing sites, like YouTube, provide an outsized choice of social media tools that area unit optimized for viewing, sharing, and embedding digital media content on the net. They conjointly give options that area unit usually found on different varieties of social media sites, like profiles, connections, comments, and personal electronic messaging. Most media-sharing sites area unit straightforward to use, give free basic accounts, and area unit accessible from each desktop and mobile devices.

Wikis area unit public forum websites that includes text and transmission content which will be altered by users. “Wiki” may be a Hawaiian sense “quick,” that refers to the speed with that data on a wiki are often accessed, added, edited, or deleted.

APIs to access Twitter information are often classified into 2 varieties supported their style and access method:

- REST genus APIs area unit supported the remainder design currently popularly used for coming up with net genus APIs. These genus APIs use the pull strategy for information retrieval. to gather data a user should expressly request it.
- Streaming genus APIs provides never-ending stream of public data from Twitter. These genus APIs use the push strategy for information retrieval. Once asking for data is formed, the Streaming genus APIs give never-ending stream of updates with no any input from the user.

They have totally different capabilities and limitations with reference to what and the way a lot of data are often retrieved.

The Streaming API has 3 varieties of endpoints:

- Public streams: These area unit streams containing the general public tweets on Twitter.
- User streams: These area unit single-user streams, with to any or all the Tweets of a user.
- website streams: These area unit multi-user streams and meant for applications that access

Tweets from multiple users requests to the genus APIs contain parameters which may embody hash tags, keywords, geographic regions, and Twitter user IDs.

IV. Healthcare and Social Media

The healthcare industry is dynamic with unimaginable speed, and one amongst the main contributors to the present modification is that the dramatic upsurge in aid communication brought on by social media. aid Social Media Analytics will bring unimaginable price by segmenting, analyzing and curating on-line aid discussions to answer your distinctive queries and wishes. Learn a lot of regarding aid Analytics. Today, at increasing rates, patients are authorized patients. they require to be concerned and to participate. However, the aid content they notice on-line isn't continuously trustworthy. Healthcare suppliers are required to guide their patients to quality content, to minister data and to be a sure supply of knowledge on-line.

Twitter's simplicity of useful style, speed of delivery and skill to attach 2 or a lot of individuals round the world provides a robust suggests that of communication, idea-sharing and collaboration. There's efficiency within the ability to burst out one hundred forty characters, as well as a shortened URI. as an example, doctors and nurses share medical data, usually as short bursts of information (lab values, conditions, orders, etc.).

V. About Swine Flue

The Swine influenza, additionally referred to as pig grippe, swine flu, hog contagion and pig contagion, is Associate in Nursing infection caused by anyone of many styles of influenza viruses. {swine grippe|swine flu|influenza|flu|grippe} virus (SIV) or swine-origin grippe virus (S-OIV) is any strain of the influenza family of viruses that's finish emicinpigs. As of 2009, the far-famed SIV strains embody grippe C and therefore the subtypes of grippe A called H1N1, H1N2, H2N1, H3N1, H3N2, andH2N3. influenza virus is common throughout pig populations worldwide. Transmission of the virus from pigs to humans isn't common and doesn't continuously result in human contagion, usually ensuing solely within the production of antibodies within the blood. If transmission will cause human contagion, it's referred to as zoon vellication flu. individuals with regular exposure to pigs ar at inflated risk of flu infection.

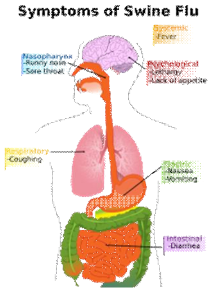
Symptoms of Swine Flu
Systemic -Fever
Nasopharynx -Runny nose -Sore throat
Psychological -Lethargy -Lack of appetite
Respiratory -Coughing
Gastric -Nausea -Vomiting
Intestinal -Diarrhea

Fig 2

VI. Machine Learning Algorithms

Machine learning techniques classified into two basic techniques as defined below.

Machine learning techniques classified into 2 basic techniques as outlined below.

Supervised learning:-

The main task here is to make a classifier. The classifier desires coaching examples which may be labelled manually or obtained from user generated user labeled on-line supply. Most used supervised algorithms are Support Vector Machines (SVM), Naive classifier and Multinomial Naive Bayes. it's been shown that supervised Techniques outdo unattended techniques in performance (Pang et al, 2002).In supervised learning the coaching knowledge set's is thought means supply that classification/regression resolution is already outlined means the classification and regression of information is properly outlined is named coaching data set's. In supervised learning the such coaching knowledge set's is thought. In supervised learning coaching knowledge set's is employed for machine learning in computing.

Example of supervised learning ways is simply like Perception, LDA, SVMs, linear/ridge/kernel ridge regression. In supervised learning the 2 major step's is employed that is "Training step", "Prediction step". In coaching step perceive regarding classifier and regress-or from coaching knowledge set, prediction step assign category labels and useful price to check knowledge.

Unsupervised learning:-

In unattended learning coaching knowledge set's isn't used for learning means the info supply that classification/ regression solutions isn't predefined, In unattended learning the info cluster and dimension reduction is includes. we tend to merely say that coaching while not teacher is named unattended learning, here teacher suggests that the coaching knowledge set's.

VII. Latent Dirichlet allocation Model

In natural language processing, Latent Dirichlet allocation (LDA) is may be a generative model that permits sets of observations to be explained by unobserved teams that designate why some components of the info are similar. as an example, if observations are words collected into documents, it posits every document may be a mixture of a little range of topics which each word's creation is as a result of one amongst the document's topics. There are several applications of LDA in many downside domains like document modeling, document classification, and cooperative filtering[18].

LDA Model

With plate notation, the dependencies among the numerous variables will be captured briefly. The boxes are "plates" representing replicates. The outer plate represents documents, whereas the inner plate represents the recurrent selection of topics and words inside a document. M denotes the amount of documents, N the amount of words in a very document. Thus:

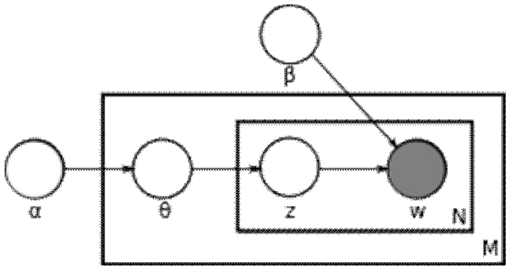

Fig 3.The graphical model for latent Dirichlet allocation

α is the parameter of the Dirichlet prior on the per-document topic distributions,
β is the parameter of the Dirichlet prior on the per-topic word distribution
θi is the topic distribution for document *i*
$\varphi_k$is the word distribution for topic *k*
$Z_{ij}$is the topic for the *j*th word in document *i*, and
$W_{ij}$is the specific word.

Tweets

C Supervised Learning

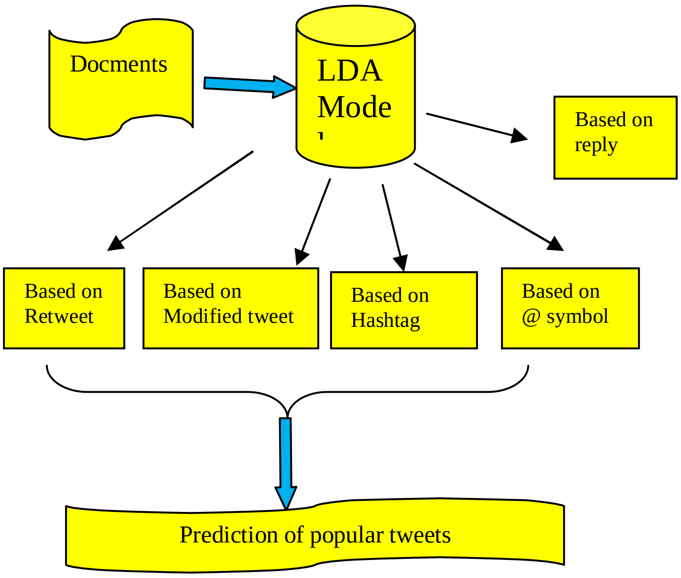

Fig 4.Tweet Classification & Prediction Using LDA Topic Modeling

Formal Classification and prediction Definition:

- Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.
- Such analysis can help to provide us with a better understanding of the data at large.
- classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions.

Classification

- The goal of data classification is to organize and categorize data in distinct classes.
- A model is first created based on the data distribution.
- The model is then used to classify new data.
- Given the model, a class can be predicted for new data.
- In general way of saying classification is for discrete and nominal values.

Prediction

- The goal of prediction is to forecast or deduce the value of an attribute based on values of other attributes.
- A model is first created based on the data distribution.
- The model is then used to predict future or unknown values.

Summarization of Classification and Prediction:

- If forecasting **discrete** value (Classification)
- If forecasting **continuous** value (Prediction)

Tasks using training data:

- Classification of tweets
- Predicting popular tweets

For classification tweets are considered as documents and then using supervised learning (LDA) we classified tweets using RT(Re tweet), MT(Modified tweet),Tweets based on @ symbol, Tweets based on #and based on reply. Then we are finding the tweet volume in each classification. To find popular messages (Prediction), we consider the number of times a message has been re tweeted.

VIII. Conclusion and Future Work

When used showing wisdom and reasonably, social media sites and platforms provide the potential to push individual and public health, still as skilled development and advancement. during this paper, 1st tweets are retrieved from twitter.com by twitter API(training set), then supervised learning methodology referred to as Dirichlet Allocation (LDA)used and mistreatment this classifier ,we have to contemplate every tweet as a document and mistreatment LDA totally different topics are elect from the document so different words associated with the topics are retrieved and appointed to corresponding topics(classification) then we tend to value however well this classification is i.e., do the topics correspond well to the tweets and that we expected the tweets that are common mistreatment re tweets..

References

1.Harris JK, Mueller NL, Snider D, Haire-Joshu D: "Local health department use of twitter to disseminate diabetes information", United States.2013.
2.Chew C, Eysenbach G: "Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak",PLoS ONE 2010, 5(11):e14.
3.Aramaki E, Maskawa S, Morita M: " Twitter catches the flu: detecting influenza epidemics using Twitter". In Proceedings of the Conference on Empirical Methods in Natural Language Processing; 27-31 July 2011. Edinburgh, Stroudsburg, PA: Association for Computational Linguistics; 2011.
4. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL: Identifying health-related topics on Twitter. In Proceedings of the Forth International Conference on Computing, Behavioral-Cultural Modeling and Prediction; 29-31 March; College Park, MD. Edited by Salerno J, Yang JS, Nau D, Chai S-K. New York: Springer Berlin Heidelberg; 2011:18-25.
5.West JH, Hall PC, Prier K, Hanson CL, Giraud-Carrier C, Neeley ES, Barnes MD: Temporal variability of problem drinking on Twitter.Open J Prev Med 2012.
6. Heaivilin N, Gerbert B, Page J, Gibbs J" Public health surveillance of dental pain via Twitter."JDentRes 2011.
7. Scanfeld D, Scanfeld V, Larson EL" Dissemination of health information through social networks: Twitter and antibiotics". Am J Infect Control 2010.
8.Hanson CL, Burton SH, Giraud-Carrier C, West JH, Barnes MD, Hansen B: "Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (Adderall) among college students",J Med Internet Res 2013.

9.Paul M, Dredze M “You are what you tweet: analyzing Twitter for public health”. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011); 17-21 July 2011.
10.Bo Pang, L.L,“ Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval”, January Volume 2 Issue 1-2, 1–94 (2008).
11.Daniel Ramage,SusanDumais,Dan, “Characterizing Microblogs with Topic Models”, Association for the Advancement of Artificial Intelligence (www.aaai.org).,2010.
12.Atingovde, I.S., Demir, E., Can, F., and Ulusoy, O. Site-based dynamic pruning for query processing in search engines. In Proc.SIGIR (Singapore, July 2008), 861-862.
13. David M. Blei,“ Surveying a suite of algorithms that offer a solution to managing large document archives”, communications of the acm, april 2012 | vol. 55 | no. 4.
14.Philipp G. SandnerAndranikTumasjan, TimmO. Sprenger and Isabell M. Welpe.Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.
15.Brendan O'Connor, RamnathBalasubramanyan,Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010.
16.D. Irani Webb, C. Pu, and K. LiS. Study of trend-stuffing on twitter through text classification. : Proceedings of 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, 2010.
17.David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3 (2003) 993-1022.
18.D. M. Blei, A. Y. Ng, and M. I. Jordan. Latentdirichlet allocation. Journal of Machine Learning Research, 3:993-1022, Mar. 2003.
19.L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis.Discovering geographical topics in the twitter stream. In Proceedings of the 21st international conference on World WideWeb, 2012.
20. D. Kim, S. Kim, and A. H. Oh. Dirichlet process with mixed random measures: A nonparametric topic model for labeled data. In Proceedings of the International Conference on Machine Learning,2012.
21. Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In Proceedings of the International Conference on World Wide Web, 2008.
22. X. Ni, J.-T. Sun, J. Hu, and Z. Chen.Mining multilingual topics from Wikipedia. In Proceedings of the 18th International Conference on World Wide Web, 2009.
23. C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In IEEE Conference on Computer on Vision and Pattern Recognition, 2009.
24.D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.